Neural Network-based Approaches for Vocal Track Extraction: Exploring Deep Clustering, Hybrid Architectures, U-net, and UH-net Models

Author Manai Mohamed Mortadha
mannaimortadha898@gmail.com

***Abstract*—Deep neural networks have become a cornerstone in various recognition and classification tasks due to their ability to learn complex patterns from raw data. This paper explores the potential application of neural networks in the domain of vocal extraction. We investigate the utilization of neural network architectures, specifically the deep clustering model based on recurrent neural networks (RNNs) and the U-net model based on convolutional neural networks (CNNs), for the task of vocal track extraction. Additionally, we propose a novel hybrid approach that incorporates a pretrained RNN model to enhance the performance of the U-net model in vocal track extraction..**

***Index Terms*—Deep neural networks, vocal extraction, recognition, classification, complex patterns, raw data, deep clustering model, recurrent neural networks (RNNs), U-net model, convolutional neural networks (CNNs), hybrid approach, pretrained RNN model, separation accuracy, spectral features, temporal context, audio source separation, separation quality, perceptual accuracy.**

I. Introduction

Vocal track extraction, a crucial component of Music Information Retrieval (MIR), involves the isolation of vocal tracks from audio files. This task finds application in diverse fields, including singer identification and lyrics transcription. Extensive efforts have been devoted to addressing this challenge, yielding numerous impactful solutions.

A prevalent strategy for vocal track extraction involves adapting techniques employed in semantic segmentation, a task that assigns class labels to individual pixels in images. This approach draws on established models for semantic segmentation, such as the widely recognized U-net model. By employing this method, vocal track masks are directly generated from audio feature maps.

Alternatively, the application of deep clustering models has gained traction. In contrast to generating vocal track masks, deep clustering models yield embedding vectors for time-frequency (T-F) bins in mel-spectrograms. In the subsequent phase, unsupervised techniques like k-means are employed to distinguish vocal T-F bins from background T-F bins.

In this project, our initial focus centers on the implementation of the models outlined in [1] and [2]. Subsequently, we endeavor to devise a novel hybrid model that amalgamates elements from both methods. A key objective is to assess the performance of the proposed hybrid model against the backdrop of traditional models. Through these endeavors, we aim to contribute to the advancement of vocal track extraction techniques in the realm of audio signal processing.

II. Related Work

A. Convolutional Neural Network and Semantic Segmentations

The inception of Convolutional Neural Networks (CNNs) stemmed from their initial role as feature extractors in image processing tasks. Over time, their utility has extended to encompass various feature recognition endeavors, including semantic parsing and audio feature extraction. CNNs have the capacity to extract hierarchical features from input signals by employing convolutional kernels.

An early breakthrough in deploying Deep Neural Networks (DNNs) for semantic segmentation tasks was marked by the advent of Fully Convolutional Networks (FCNs). The FCN architecture engages in the extraction of intricate features from input images utilizing conventional CNN structures, such as the widely employed VGG16 network. To translate the compact feature maps back into segmentation outcomes, FCN integrates an upsampling mechanism. This process facilitates the restoration of higher-resolution segmentation results from the downscaled feature maps.

In this evolutionary trajectory, CNNs have evolved from their origins in image analysis to serve as powerful tools for a broader array of feature recognition tasks. FCNs, as a prominent example, underscore the adaptability of CNN architectures in addressing complex challenges like semantic segmentation through the adept integration of upsampling strategies.

The innovative approach pioneered by Fully Convolutional Networks (FCNs) involves crafting segmentation masks that mirror the dimensions of the input, transforming the intricate task of segmentation into a more manageable pixel-level classification problem. By aligning the segmentation output with the original input's layout, FCNs redefine the challenge in terms of assigning class labels to individual pixels, allowing for a more granular analysis.

To bolster the fidelity of information during the crucial upsampling phase, FCNs introduce a strategic fusion of the feature map with outputs from intermediate layers, such as pool4 or pool3 [3]. This amalgamation ensures that high-level contextual information is retained during the transition from the downscaled feature maps to the final segmentation mask. This context-aware strategy plays a vital role in maintaining the integrity of the original data during the upsampling process.

However, despite its innovative approach, the FCN model exhibits limitations when applied to certain scenarios, such as music source separation. One significant drawback lies in its struggle to preserve intricate details in proximity to separation boundaries. This shortcoming impedes its suitability for tasks that demand a high degree of precision around such boundaries, as often encountered in audio source separation.

Enter the U-net architecture, an influential advancement built upon the foundation of FCN models. U-net, initially renowned for its prowess in biomedical image segmentation, offers refinements that bolster its performance in challenging scenarios. A key distinction is the augmentation of the channel count within the upsampling convolutional layers. This strategic enhancement enables the propagation of richer contextual information to the higher-resolution layers, facilitating a more comprehensive understanding of the input.

Another noteworthy innovation in the U-net model is the assignment of increased loss weights to boundary pixels. This strategic adjustment effectively emphasizes the accurate depiction of boundary regions, leading to the generation of more precise and well-defined masks. This tactic is particularly beneficial when handling intricate boundaries, such as those prevalent in medical imaging.

The U-net model's adaptability and robustness have extended beyond its original application domain. Its ability to perform well with limited datasets underscores its potential for various scenarios, including music source separation tasks. By capitalizing on its enhanced contextual understanding and improved boundary preservation, the U-net model opens new avenues for addressing challenges in audio signal processing, demonstrating its potential as a versatile tool in diverse fields.

B. Recurrent Neural Network

The Recurrent Neural Network (RNN) architecture has been meticulously crafted to tackle the intricate challenges posed by sequential input data. RNN models have found widespread utility across diverse domains, prominently including natural language processing and the intricate landscape of audio signal modeling. It is noteworthy that the Gated Recurrent Unit (GRU), a remarkable innovation within the RNN framework, has garnered considerable attention for its prowess. Sharing commonalities with the Long Short-Term Memory (LSTM) architecture, GRUs exhibit a distinctive gating mechanism. However, their distinctive strength lies in their parsimonious parameterization, achieved by obviating the need for an output gate component.

Venturing into the realm of deep clustering models, the bedrock of the feature extraction process consists of a four-layer Bidirectional Long Short-Term Memory (Bi-LSTM) network. This intricate network structure serves as a potent catalyst, orchestrating the metamorphosis of input mel-spectrograms into high-dimensional embedding vectors. A seminal work highlighted in reference [2] embarks on a groundbreaking journey by synergistically amalgamating conventional neural networks with the deep clustering paradigm. The overarching objective is to elicit enhancements in the overall model performance and efficacy.

In this endeavor, our pursuit is to meticulously replicate the architectural blueprint outlined in the aforementioned study. However, what sets our contribution apart is the calculated substitution of the traditionally employed Bi-LSTM layers with their Bidirectional Gated Recurrent Unit (Bi-GRU) counterparts. This strategic substitution is not only tactful in simplifying the training process but also harbors the latent potential to unlock heightened model efficiency and refined performance benchmarks. The incorporation of Bi-GRUs infuses a new dimension into the design space, offering tantalizing prospects for augmenting model efficiency, while ensuring that the integrity of the model's predictive prowess remains undeterred.

III. DATA PREPOSSESSING

A. DSD100 Dataset

Renowned within the realm of Music Information Retrieval (MIR) tasks, the DSD100 dataset stands as a prominent benchmark. Comprising a curated collection of 100 complete music tracks, this dataset is accompanied by their individual isolated tracks. To enhance the diversity and comprehensiveness of our training data, a nuanced approach is undertaken. While the original music tracks are meticulously preserved, a strategy of deliberate amalgamation is employed. This involves the randomized fusion of vocal tracks sourced from the DSD100 dataset with assorted instrumental tracks. The outcome of this endeavor is the generation of an expanded corpus of audio files that exhibit a rich interplay of vocal and instrumental elements, elevating the quality and diversity of the training data pool.

B. Multiframe

Employing a sophisticated approach detailed in reference [4], the multiframe strategy serves as a catalyst in both augmenting training data and streamlining the training process. At the outset, a meticulous process ensues, involving the excision of silence segments from both vocal tracks and their corresponding segments in mixed music tracks. Subsequently, a transformation transpires, converting the amalgamated mixed music tracks and veritable ground truth voice tracks into the realm of log-scaled mel-spectrograms. These transformations, calibrated with 128 mel bands and 16000 sample rate, yield compact mel-spectrogram chunks, each encapsulating 128x128 feature maps. These chunks, averaging approximately one-second audio sequences, unravel the temporal and spectral intricacies.

It's imperative to observe the spatial reorientation within the mask's representation. In this schema, the x-axis within the mask pertains to features, while the y-axis corresponds to time. This alignment is paramount, particularly as all spectrograms undergo a transposition process to match this conceptual framework.

In the pursuit of vocal track extraction from music compositions, a dual-phase strategy unfurls. Beginning with the conversion of log-scaled mel-spectrograms into power-scaled equivalents, a pivotal step unfolds. This involves the multiplication of the spectrogram by a meticulously generated filter

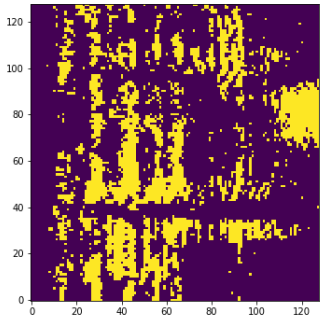

Fig. 1. Mask for vocal track : Binary Mask

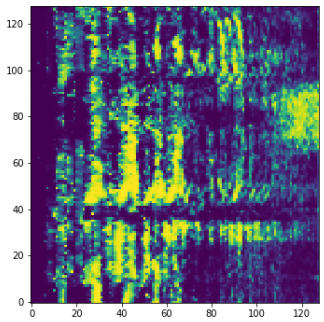

Fig. 2. Mask for vocal track : Proportion Mask

mask, forged by the underpinning models. The ensuing stage navigates the reversion of the vocal track spectrogram into its auditory counterpart.

In the context of each music-vocal chunk pair, the strategic design entails the construction of dual training target masks. The first, a binary vocal mask, materializes through a juxtaposition of the power between vocal and background spectrograms. This discerning juxtaposition demarcates the temporal-frequency bins, demarcating their allegiance to either vocal or background sources. This binary mask, carefully calibrated, serves as a foundational component for training clustering models.

Concurrently, a secondary target mask emerges—the proportion mask—relinquishing a distinct perspective. This mask encapsulates the dominance exerted by the vocal source across the mel-spectrogram's time-frequency bins. This nuanced representation takes center stage during the training of the U-net model, orchestrating its learning dynamics with precision. The distinct roles and implications of these dual masks embody the multifaceted nature of the training process, encapsulating the intricate interplay between clustering models and the U-net architecture.

C. Vocal Track Recovering

In the absence of preserved phase information during audio data processing, the restoration of the vocal track from mel-spectrograms necessitates the integration of the Griffin Lim algorithm [5]. This algorithm, operating through iterative cycles, orchestrates the simulation of phase information to facilitate the meticulous reconstruction of audio signals.

IV. MODEL DESCRIPTION

A. Deep Clustering Model

The structural blueprint of the deep clustering model mirrors the framework established in prior literature [2]. To tailor the model to our context, we undergo strategic modifications, notably by downsizing the input mel-spectrogram dimensions from 150x150 to 128x128. Additionally, a pivotal adaptation entails the substitution of the LSTM layer with a GRU layer. Within this revamped configuration, a four-layer Bi-GRU network orchestrates the assignment of D-dimensional feature vectors to each time-frequency (T-F) bin. Notably, the dimensionality D is conservatively set to 20, a value in consonance with recommendations outlined in [2].

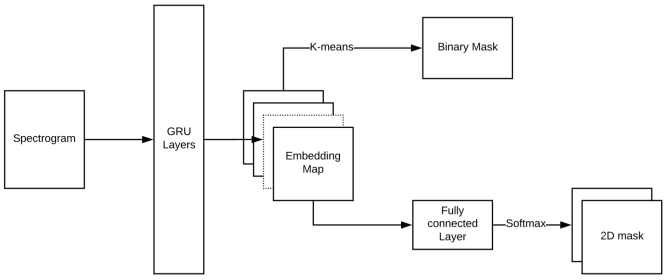

Fig. 3. The hybrid clustering model

Venturing further, we delve into the intricate terrain of the hybrid network, detailed in [2]. This novel structure is anchored in the core architecture of the deep clustering model while introducing a novel facet—a supplementary head that generates a mask exhibiting softmax activation. In the intricate dance of model computation, this supplementary mask embodies a two-dimensional vector assigned to each T-F bin. This vector encapsulates the dynamic interplay between vocal and background sources, effectively quantifying their respective contributions to the aggregate power composition.

B. U-net based model

In the initial phase of our study, we undertake the replication of the classical U-net architecture elucidated in reference [3]. The U-net model, a cornerstone of our exploration, is characterized by the amalgamation of four downsampling layers and an equivalent number of upsampling layers. While originally conceived as a semantic segmentation model within the domain of image processing, our observations unveil an intriguing revelation. Namely, the U-net model seamlessly transitions to the arena of vocal extraction, where its intrinsic capabilities aptly apply. This seamless translation of utility from image semantics to the intricacies of vocal extraction underscores the model's versatility and adaptability across disparate domains.

In contrast to the swift convergence and favorable generalization exhibited by the deep clustering model, a notable characteristic of the U-net model emerges—prolonged training times coupled with a propensity to succumb to overfitting

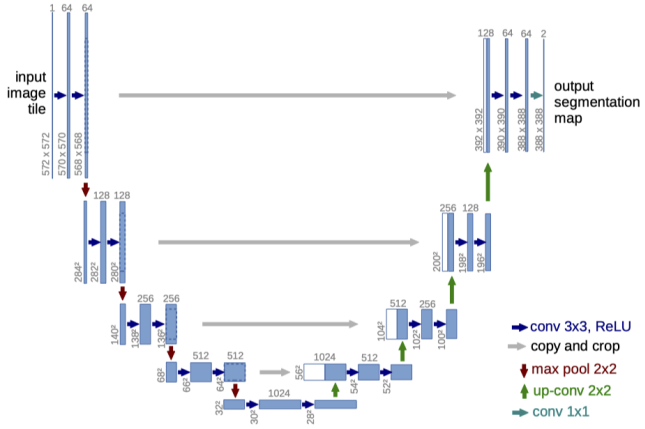

Fig. 4. U-net model[3]

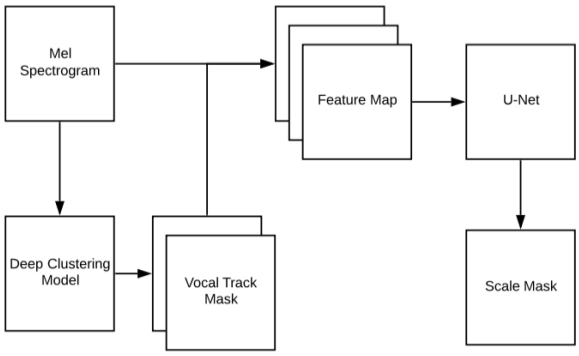

Fig. 5. he hybrid U-net model

when presented with the current training dataset. Responding to this multifaceted challenge, we embark on an innovative exploration—a symbiotic fusion that intertwines the strengths of the deep clustering model and the U-net architecture. Our motivation lies in investigating whether this confluence can catalyze transformative improvements.

As we delve into the architectural subtleties, the original U-net model, originally harnessed as a semantic segmentation powerhouse for image analysis, merits our attention. In its native configuration, the U-net model operates within the contours of a single input channel, dedicated to the representation of log mel-spectrograms. However, our journey into innovation compels us to chart an unconventional course. The evolved U-net variant embraces a novel dimension, as it undertakes a harmonious dual training regimen alongside the deep clustering model. This harmonization transpires through the integration of the unadulterated spectrogram with the outcome of the deep clustering model—manifested as a softmax mask. This judicious melding fabricates a dynamic trichromatic feature map, meticulously curated to serve as the input palette for the U-net's computational domain.

This revolutionary amalgamation, which begets the hybrid U-net model, ushers in a promising era of exploration. Empirical validation, a harbinger of transformative insights, attests to the model's performance supremacy vis-à-vis its conventional predecessor. This testament to the hybrid model's ascendancy underscores its potential to overcome the challenges that challenge the standard U-net's efficacy. It is imperative to underscore that both incarnations of the U-net paradigm—the conventional and the hybrid—intake mel-spectrograms of dimensions 128x128 as input. Their shared output—proportion masks—paints a vivid picture, embodying the degree to which the vocal source wields its influence over the time-frequency bins encased within the intricate feature map. This dual journey of architectural innovation and nuanced output encapsulates our comprehensive quest to elevate vocal track extraction through the lens of neural networks.

V. EXPERIMENT

A. Training

1) Deep Clustering Models: The operational framework of deep clustering models entails a rigorous training regimen. Employing a batch size of 32, the rmsprop algorithm stands as the optimizer of choice, steering the model towards convergence. Our training initiation focuses exclusively on the embedding component of the model. Central to our training objective is the pursuit of convergence between the affinity matrix derived from the model's generated embedding output and the corresponding binary mask.

The binary mask, denoted as Y, is reshaped into a matrix $Y \in \mathbb{R}^{TF\times 1}$, where T signifies the number of frames and F signifies the feature dimensions within the feature map. The output of the clustering model, embodied in a matrix $V \in \mathbb{R}^{TF\times D}$, takes its place as a fundamental participant. Here, D materializes as the dimensional expanse of the embedding dimension. Anchored in this context, the loss function materializes, encapsulated by the expression

$$L = \|VV^T - YY^T\|_F^2 \tag{1}$$

However, a noteworthy consideration underscores this process—precise execution entails resource-intensive matrix multiplications that demand substantial GPU memory. To circumvent these challenges and streamline computation, we pivot towards a simplified loss function representation:

$$L = ||V^TV||_F^2 + ||Y^TY||_F^2 - 2||V^TY||^{2[6]} \tag{2}$$

.

This transformation optimizes the calculation process while upholding fidelity to our objective. At its core, the deep clustering model crystallizes through a learning process that approximates the affinity matrix—an approximation derived from the deep clustering model—by minimizing the objective function.

Transitioning to the hybrid variant, the focal point shifts to the proportion-mask head. Within this architectural domain, the training mandate is distinctly delineated: facilitate the generation of a vocal source proportion mask mirroring the size of the input. The binary cross-entropy (BCE) loss emerges

as the cornerstone of our objective function, serving as a potent tool to gauge and optimize performance. This intricate interplay of loss functions and model optimization manifests the intricate choreography that propels the deep clustering and hybrid deep clustering models towards their respective performance zeniths.

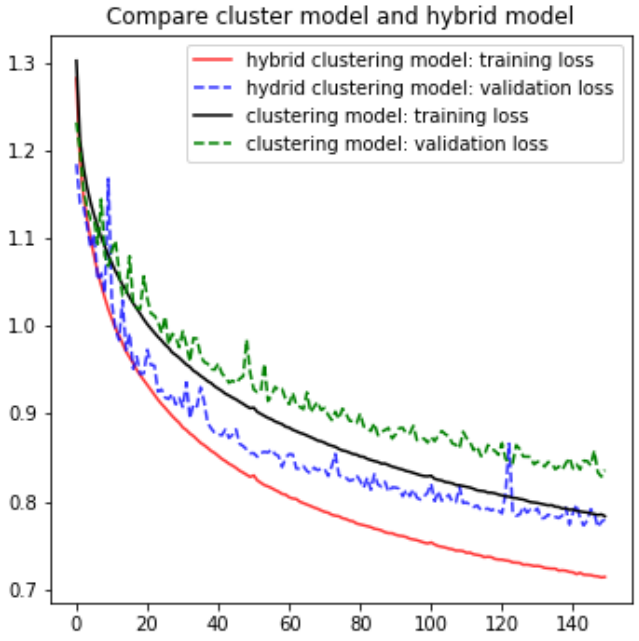

Fig. 6. Loss vs Epoch for clustering models

As shown in the Fig 5, the hybrid clustering model with an extra head has better performance. The addition of proportion-mask head speeds up the training process and also decreases the model's overfitting.

2) U-net Based Models:

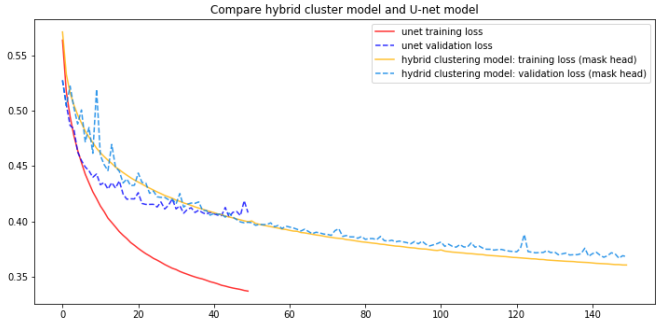

Fig. 7. Loss vs Epoch for U-net/Clustering models

The realm of U-net models unfolds with the rmsprop algorithm steering the optimization trajectory. The batch size, an intrinsic parameter of the model's optimization, is meticulously tuned to 32—a pivotal choice that orchestrates convergence dynamics. Rooted in the essence of vocal track extraction, the proportion mask stands as the designated training target, aptly lending itself to the broader orchestration of Backpropagation through Time (BPTT). In this intricate symphony of learning, the Binary Cross-Entropy (BCE) Loss plays a central role—a resonant force that facilitates the fine-tuning of the model's parameters.

The pursuit of precision prompts the strategic deployment of dropout layers, a tactical move engineered to combat the looming specter of overfitting. These layers, infused with a dropout rate of 0.5, render the model inherently more resilient by imposing a controlled measure of randomness—effectively mitigating the risk of overly specialized learning.

In the visual narrative of Fig. 4, a compelling exposition unfurls, comparing the foundational U-net model with the proportion-mask head of the deep clustering model. A discernible revelation emerges—underscored by empirical evidence—the U-net model ascends as the frontrunner, boasting an unequivocally superior performance trajectory vis-à-vis the deep clustering model. However, beneath the surface, a nuanced dichotomy comes to light. Despite its evident prowess, the traditional U-net model grapples with a cardinal challenge—its inherent limitations in generating the high-resolution delineation required to disentangle vocal sources from the backdrop of background voice within the intricate feature map.

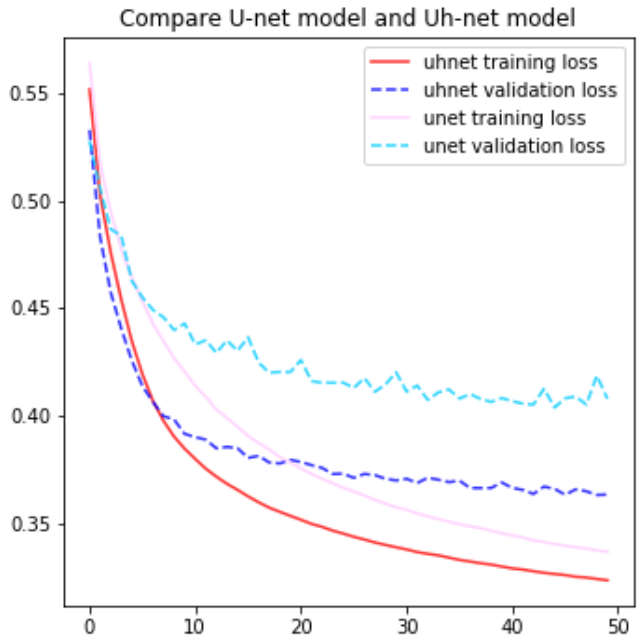

Fig. 8. Loss vs Epoch for U-net/UH-net models

This quandary paves the way for innovation—the conception of a hybrid architecture christened UH-net. This novel synthesis artfully marries the advantages intrinsic to both the U-net and deep clustering models. The UH-net's metamorphic journey charts an evolution that deftly evades the pitfalls of overfitting, seamlessly preserving the U-net's prowess while superimposing the rich insights harvested from the deep clustering model. Fig. 5 furnishes a tangible vista—depicting the trajectory of loss against epochs for both the U-net and the transformative UH-net. A palpable shift in performance transpires—the UH-net's convergence is marked by swifter

strides and a pronounced dip in validation loss, underscoring the efficacy of this fusion. This harmony of architectural synthesis unlocks a novel dimension, showcasing the relentless pursuit of excellence in the realm of vocal track extraction.

B. Results

1)U-net Based Models: The realm of U-net models unfolds with the rmsprop algorithm steering the optimization trajectory. The batch size, an intrinsic parameter of the model's optimization, is meticulously tuned to 32—a pivotal choice that orchestrates convergence dynamics. Rooted in the essence of vocal track extraction, the proportion mask stands as the designated training target, aptly lending itself to the broader orchestration of Backpropagation through Time (BPTT). In this intricate symphony of learning, the Binary Cross-Entropy (BCE) Loss plays a central role—a resonant force that facilitates the fine-tuning of the model's parameters.

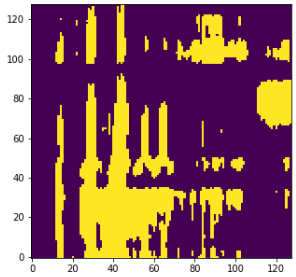

Fig. 9. Clustering Model

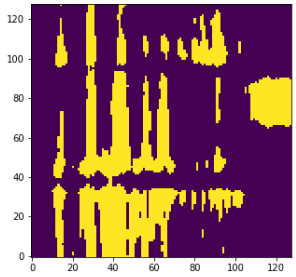

Fig. 10. Hybrid Clustering Model

The pursuit of precision prompts the strategic deployment of dropout layers, a tactical move engineered to combat the looming specter of overfitting. These layers, infused with a dropout rate of 0.5, render the model inherently more resilient by imposing a controlled measure of randomness—effectively mitigating the risk of overly specialized learning.

In the visual narrative of Fig. 4, a compelling exposition unfurls, comparing the foundational U-net model with the proportion-mask head of the deep clustering model. A discernible revelation emerges—underscored by empirical evidence—the U-net model ascends as the frontrunner, boasting an unequivocally superior performance trajectory vis-à-vis

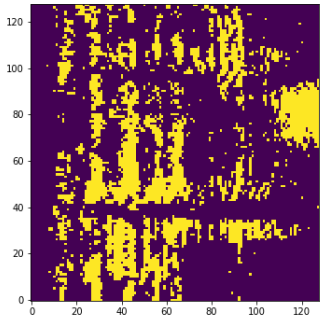

Fig. 11. Ground Truth

the deep clustering model. However, beneath the surface, a nuanced dichotomy comes to light. Despite its evident prowess, the traditional U-net model grapples with a cardinal challenge—its inherent limitations in generating the high-resolution delineation required to disentangle vocal sources from the backdrop of background voice within the intricate feature map.

This quandary paves the way for innovation—the conception of a hybrid architecture christened UH-net. This novel synthesis artfully marries the advantages intrinsic to both the U-net and deep clustering models. The UH-net's metamorphic journey charts an evolution that deftly evades the pitfalls of overfitting, seamlessly preserving the U-net's prowess while superimposing the rich insights harvested from the deep clustering model. Fig. 5 furnishes a tangible vista—depicting the trajectory of loss against epochs for both the U-net and the transformative UH-net. A palpable shift in performance transpires—the UH-net's convergence is marked by swifter strides and a pronounced dip in validation loss, underscoring the efficacy of this fusion. This harmony of architectural synthesis unlocks a novel dimension, showcasing the relentless pursuit of excellence in the realm of vocal track extraction. We calculate the IoU score on test set for both models. The IoU is calculated by the following equation:

$$IoU = \frac{\text{Area of overlap}}{\text{Area of union}} \tag{3}$$

For the basic model, the IoU accuracy is 74.78and the IoU of hybrid model is 79.44where we can see a explicit improvement.

2) U-net Based Models: A symphony of innovation reverberates within the U-net based models, culminating in the creation of vocal proportion masks that navigate the intricate terrain of source separation. The vivid tapestry of these masks materializes in Fig. 9, where a visual odyssey unfolds, showcasing the prowess of the UH-net model—a sentinel that not only preserves intricate details but also paints explicit segmentation boundaries with an artistic finesse.

The assessment of how adeptly the predicted mask aligns with the hallowed ground of ground truth calls for an objective metric. In this narrative of evaluation, the Peak Signal-to-Noise Ratio (PSNR) algorithm stands as a reliable companion, poised to measure the semblance between different images. While traditionally harnessed for image comparison, the PSNR

algorithm finds an unexpected resonance within our domain, rendering it an apt tool to gauge our model's performance.

The equation that governs this calculation reads:

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{\text{R}^2}{\text{MSE}} \right) \tag{4}$$

Where MAX denotes the maximum possible pixel value and MSE signifies the Mean Squared Error between the predicted mask and the ground truth. This formula resonates as a beacon of assessment, quantifying the harmony between prediction and reality—a testament to the model's efficacy and the precision of its vocal track extraction.

This symphony of assessment underscores the dynamic intersection of technology and artistry, as U-net based models pave the way for a harmonious marriage of vision and sound.

The UH-net has a PSNR score for 17.12, while the U-net model's baseline score is 16.24. We take these scores as a quantitative metric of performance.

3) Music Recovery Results: Embarking on a quest for holistic comparison, we navigate the realm of model outputs, each endowed with a distinct format. This variance in format prompts an ingenious transformation—a process that orchestrates the metamorphosis of masks into vocal tracks. This transformative feat unfolds through a delicate dance of multiplication, as the model's output mask entwines with the mel-spectrogram in its native power-scale guise. This union begets a transformed spectrogram—a harmonious symphony that resonates with the essence of the original vocal source. This newly transmuted spectrogram, a captivating embodiment of harmonic metamorphosis, then yields to the artistry of the Griffin-Lim algorithm—a maestro of audio reconstruction. The algorithm, characterized by its capability to simulate phase information through iterative alchemy, rekindles the vocal track from the depths of the spectrogram.

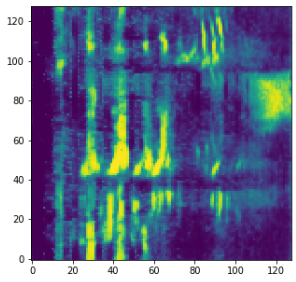

Fig. 12. UH-net Model

In our grand symphony of evaluation, full-length songs come to the forefront—a panorama punctuated by an ensemble of small frames. This partition of the musical tapestry into these frames sets the stage for a meticulous evaluation—an evaluation wherein the virtuosity of our models is spotlighted. Each frame is an individual vignette—a canvas upon which the intricate ballet of vocal extraction unfurls. Through this choreography of assessment, a panoramic narrative of model performance emerges, offering insights that transcend mere technicalities.

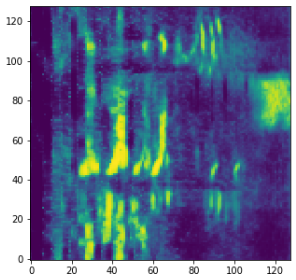

Fig. 13. U-net Model

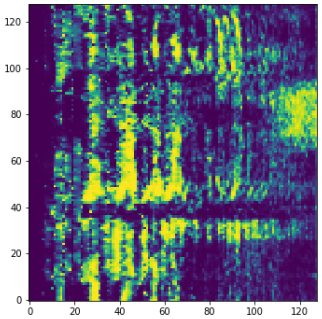

Fig. 14. Ground Truth

The symphony of music recovery results harmonizes art and science—a melodious testimony to the intersection of neural networks and the rich tapestry of music.

for each frame chunk, then we combine all results together to get the pure vocal track for the whole song. Fig 10 shows the power-scale spectrogram generated by different models for a sample one second length frame, along with the original mix and vocal tracks. One thing needs to be mentioned here is that since k-means cannot specify which T-F bin group is vocal-dominated when we implement deep clustering models, we are using the proportion-mask head output in hybrid deep clustering model to select the right group.

We calculate source to distortion (SDR) as the metric for vocal track separation model, the result is shown in the table. We test models on DSD100 test set.

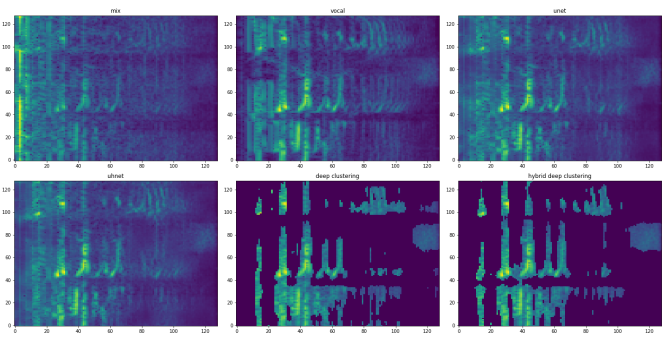

Fig. 15. Spectrogram

SDR SCORE FOR EACH MODEL	
UH-net	5.92
U-net	4.34
HDC	1.56

The landscape of model performance materializes with unmistakable clarity—our UH-net model emerges as the epitome of superior efficacy. The realm of achievement, however, unfurls amidst the horizon of potential. While the horizon is promising, the baseline gleaned from vocal extraction's annals in [2] casts a spotlight on the journey ahead. With a baseline of 6.3, the embers of optimization flicker—a testament to the ongoing quest for refinement and the boundless room for growth.

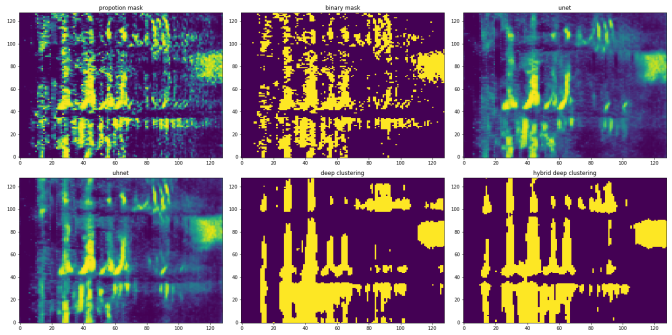

Fig. 16. Mask

The theater of testing yields an enigmatic revelation—a facet of the deep clustering model's persona that illuminates its inherent limitations. Amidst the symphony of sound, the deep clustering model might falter in the face of music void of vocal components or those with faint vocal undertones. Fig. 11 stands as a visual allegory—a window into this dynamic. This tableau of sound reveals a telling narrative—the clustering model, confined by its inherent intricacies, grapples to segregate the T-F bins in a harmonious ensemble. Contrastingly, the U-net based models, guided by their intrinsic architecture, gracefully navigate these waters, circumventing the pitfall with an elegant finesse.

As the tale unfurls, a symphony of challenges and triumphs resounds—a melody that underscores the relentless pursuit of excellence. With each stride forward, the interplay of models, nuances, and the ever-evolving soundscape harmonizes to reshape the landscape of vocal track extraction.

VI. CONCLUSION

In this paper, we delved deeply into the realm of vocal track extraction, unravelling a tapestry woven from four distinct models. These models, bearing the imprint of innovation, are borne of two principal theoretical foundations, each encapsulating unique paradigms. The first cornerstone rests upon the bedrock of deep clustering—a symphony orchestrated by embedding and the symposium of unsupervised learning. The second, steeped in the philosophy of semantic segmentation, transposes the intricacies of music source separation onto the canvas of image processing.

From this theoretical landscape, the UH-net model emerges as the magnum opus—an amalgamation of ingenuity that elegantly surpasses its counterparts. The introduction of an auxiliary clustering head breathes life into the U-net architecture—catapulting the model's training trajectory into a dimension of efficiency, while concurrently amplifying its fidelity to truth. This synergy underscores a pivotal facet of this study—enhancing the delicate equilibrium of swiftness and precision.

However, our voyage through the crucible of quantitative analysis reveals that, in the grand scheme of model accuracy, uncharted territories beckon. A panoramic vista unfurls, wherein the expansion of dataset horizons and the augmentation of resolution augur avenues for advancement. Further refinement, a symphony of architectural finesse, perpetuates the evolutionary pulse of these models—ushering forth an era where precision and innovation continue to harmonize.

APPENDIX

The project's code part and sample vocal extraction results are uploaded to GitHub , the repository's link is : https://github.com/MortadhaMannai/VOCAL-TRACK-EXTRACTION-USING-NEURAL-NETWORKS

ACKNOWLEDGMENT

we wish to convey our deepest appreciation to Professor Bassem Ben Hamed from the Mathematics and BI Department at the National School of Electronics and Telecommunications of Sfax and the Co-fondateur of DataCamp Training. His astute guidance, invaluable suggestions, and unswerving support have significantly enriched the fabric of this paper, imbuing it with an undeniable brilliance that emanates from his profound expertise.

REFERENCES

[1] Jansson, Andreas, et al. "Singing voice separation with deep U-Net convolutional networks." (2017).
[2] Luo, Yi, et al. "Deep clustering and conventional networks for music separation: Stronger together." 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017.
[3] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015.
[4] Perraudin, Nathanal, Peter Balazs, and Peter L. Sndergaard. "A fast Griffin-Lim algorithm." 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. IEEE, 2013.
[5] Wang, Zhong-Qiu, Jonathan Le Roux, and John R. Hershey. "Alternative objective functions for deep clustering." 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018.